INFANT BEHAVIOR AND DEVELOPMENT 7, 361–381 (1984)

The Intermodal Representation of Speech in Infants*

PATRICIA K. KUHL AND ANDREW N. MELTZOFF
University of Washington

Infants' abilities to detect auditory-visual correspondences for speech were tested in two experiments. Infants were shown two visual images side-by-side of a talker articulating, in synchrony, two different vowel sounds, while a sound matching one of the two vowels was auditorially presented. Infants' visual fixations to the two faces were video-recorded and scored by an independent observer who could neither see the faces nor hear the sounds. The results of Experiment 1 showed that the auditory stimulus systematically influenced infants' visual fixations. Infants looked longer at the face that matched the sound. In Experiment 2, the same visual stimuli were presented, but the auditory stimuli were altered so as to remove the spectral information contained in the vowels while preserving their temporal characteristics. Performance fell to chance. Taken together, the experiments suggest that infants recognize the correspondences between speech information presented auditorially and visually, and moreover, that this correspondence is based on the spectral information contained in the speech sounds. This suggests that infants represent speech information intermodally.

speech perception intermodal representation faces auditory stimuli
visual fixation cross-modal matching

Speech perception has traditionally been studied almost exclusively as an auditory phenomenon. Yet conversational speech is often produced by a talker we can both see and hear. What effect does visual information have on the perception of speech?

Research on adults suggests that the effect of the visual modality is considerable (Erber, 1975). Sighted people, both hearing-impaired and normal, demonstrate the ability to "lipread"—to derive linguistic information by watching a talker's mouth movements. While the clinical impact of this ability has long been recognized (Johnson, 1775), the importance of lipreading phenomena for theories of speech perception has often been overlooked.

Normal adults use visual information when they listen to speech in a noisy environment (Erber, 1969; Ewertsen & Birk-Nielsen, 1971; O'Neill, 1954; Sumby & Pollack, 1954), or when the auditory information in speech is degraded by filtering (Binnie, Montgomery, & Jackson, 1974; Sanders & Good-

*Portions of this research were presented at the 91st meeting of the Acoustical Society of America (Chicago, 1981) and a brief report appeared in Kuhl and Meltzoff (1982). The research was supported by a grant from the National Science Foundation (BNS 8103581) to P.K.K.; preparation of the manuscript was supported by this grant and by grants from the National Science Foundation to A.N.M. (BNS 8309224) and the Spencer Foundation to A.N.M. and P.K.K. Requests for reprints should be sent to Patricia K. Kuhl, Department of Speech and Hearing Sciences, University of Washington, Seattle, WA 98195.

rich, 1971). Listeners also rely on visual information when the speech of one talker is presented against a background of other talkers (Summerfield, 1979), as in a "cocktail party" situation.

The data show that there is a substantial contribution of visual information in each of these circumstances. For example, the Sumby and Pollack (1954) study found that seeing the face of a talker whose speech was presented in noise was equivalent to increasing the signal by 15–20 decibels. More recently, Grant, Ardell, Kuhl, & Sparks (in press) presented an isolated pure tone whose amplitude and pitch followed those of the fundamental frequency of a talker who was reading from text. The pure tone preserved the rhythm and stress pattern of speech and was perceived as voice-like. When the tone was presented in the absence of seeing the talker's face, no syllables, words, or phrases could be identified. However, when subjects saw the talker while listening to the tone, the spoken text became 80% intelligible. Thus, while visual information is not essential to the perception of speech (blind adults perceive speech normally), numerous studies demonstrate that vision can be and is used to derive information about speech under certain circumstances.

What kind of linguistic information does vision provide? Watching a talker's mouth movements provides two kinds of information: prosodic and phonetic. While careful work on the prosodic information available through lipreading has not been done, it appears that prosodic information in the form of syllable timing and rhythm is provided by the temporal sequences of mouth openings and closings. Since consonants are produced with a relatively constricted vocal tract and vowels with a relatively open vocal tract, the open-close cycle that results when consonant-vowel syllables are combined provides a rough visual marking of the boundaries of syllables (Erber, 1977). Syllabification is essential to the perception of stress and rhythm in speech.

Studies have also demonstrated that the visual channel provides phonetic information. The studies show that vision provides information about the "place of articulation" feature, one that distinguishes sounds like */b/, /d/,* and */g/*. These phonetic units differ in the location of the primary constriction in the mouth (*/b/* = the lips; */d/* = the alveolar ridge; */g/* = the velum), and these differences are detectable by eye when watching mouth movements. Other speech features, however, are not visually distinct. Vision does not provide information about sounds that differ in the "manner of articulation" when they occur at the same place of articulation (such as the sounds */p/, /b/,* and */m/*). These articulations involve the same primary constriction and thus are not visually distinguishable.

The availability of featural information through the auditory and visual modalities forms an interesting complimentary relationship. Research shows that speech information in the two modalities is differentially resistant to the effects of degradation, such that information that is subject to degradation in one modality is available in the other. As we suggested, place information is

available visually, while manner information is generally not. Conversely, while both place and manner information are normally available through the auditory channel, place information is much more subject to the effects of noise and/or filtering. Relatively slight increases in the background noise destroy place information auditorially, while manner information is highly resistant to these effects (Miller & Nicely, 1955). The implication is that the auditory fragility of place information can be effectively counteracted by its availability through the visual modality.

This is best illustrated in research on severely hearing-impaired listeners. Studies show that these individuals are capable of perceiving manner features such as "voicing" and "nasality" by ear, but that they are unable to distinguish the place feature auditorially (Erber, 1972). Specifically, the results demonstrate that severely impaired individuals can still distinguish voiced consonants (*/b, d, g/*) from voiceless consonants (*/p, t, k/*). However, the severely hearing-impaired listener cannot perceive place information auditorially, and thus the three place categories, involving bilabial (*/p, b, m/*), alveolar (*/t, d, n/*), and velar (*/k, g/*) sounds are not distinguished.

In essence, this means that when severely hearing-impaired listeners are auditorially presented with the consonant */p/* (a voiceless, non-nasal, bilabial sound) they correctly perceive the voiceless and non-nasal features, but are unable to perceive its place of articulation. They are therefore unable to identify the sound as */p/* as opposed to */t/* or */k/*. However, when these same hearing-impaired listeners are tested under conditions in which they watch the talker speak, the place feature can be identified by eye and performance on the consonant identification task is near perfect (Erber, 1972). Apparently, information leading to the identification of phonetic features can be independently extracted by the two modalities and then combined.

Current models of the speech perception process cannot account for the integration of auditory and visual information in the perception of speech. Nonetheless, such findings (see also McGurk & MacDonald, 1976; Summerfield, 1979) raise central questions about the representation of speech. At a minimum, they suggest that speech perception is not solely the province of audition. Rather, they suggest that information about speech can be picked up by different modalities and integrated by perceptual mechanisms to form a phenomenally unified phonetic percept. While this much is demonstrated, the manner in which the different modalities interact and the form in which speech sounds are represented to allow this interaction (in articulatory terms, auditory terms, or some more abstract phonetic representation that is not exclusively auditory or articulatory) is still unknown.

From a theoretical standpoint, the fundamental importance of these classic lipreading studies is that they suggest that, for adults, speech information derived from the visual modality can substitute for speech information derived from the auditory modality. They suggest, for example, that the visual

perception of place of articulation (conveyed by the configuration of the lips, tongue, and jaw) can substitute for the auditory perception of place of articulation (conveyed by the configuration of formant frequencies). The question is, how is information that is processed by two separate modalities—audition and vision—equated in speech perception?

One possibility is that adults, through a protracted period during which they both watch and listen to others speak, learn to associate the auditory and visual concomitants of speech. That is, they learn that an auditory /b/ is accompanied by a visible closure of the lips, and so on. If this were the case, then young infants who have not had a long period during which to learn the association between the auditory and visual concomitants of speech would be unable to relate them.

The aim of this experiment was to begin to explore the development of auditory-visual speech perception. We asked whether 4-month-old infants recognized that sounds of a particular type were emitted by mouths moving in a particular way. Specifically, our question was whether infants could detect cross-modal correspondences for speech presented to the auditory and visual modalities.

The problem was posed by showing infants two faces, side-by-side, articulating two different vowel sounds. A sound track matching one of the faces was auditorially presented. We hypothesized that the auditorially presented signal would systematically influence infants' visual preferences. Specifically, we suggested that if infants recognized the correspondence between articulatory gestures and their auditory consequences, they would look longer at the face whose articulatory movements matched the sound presented.

An initial report of the cross-modal speech perception effect was provided by Kuhl and Meltzoff (1982). The purpose of the present paper is to provide full methodological details of the stimulus preparation, experimental procedure, and results, along with a more complete analysis of the theoretical implications of these findings.

EXPERIMENT I

METHODS

Subjects

Thirty-two infants served as subjects. They had no known visual or auditory abnormalities. They ranged in age from 18.0 weeks to 20.1 weeks ($M = 19.3$). Participation in the experiment was solicited by a letter that was sent to the parents of newborns in the Seattle area. Interested parents returned a postcard, which provided details regarding birth and family medical history. Infants who were preterm, low birth weight, or otherwise at-risk for normal development were not tested. An additional 10 infants failed to complete testing due to crying (5), falling asleep (2), or equipment failure (3). Parents were paid $3 for their participation in the study.

Stimuli

An Auditory-Visual Speech Perception technique (AVSP) was developed to present stimuli to the infants. Using this technique, infants are shown two filmed images, side-by-side, of a female talker articulating in synchrony two different vowels. The sound track corresponding to one of these vowels is presented through a loudspeaker directly behind the screen and midway between the facial images. In this study we wanted to examine infants' knowledge that particular types of speech sounds are produced by mouths moving in particular ways. In order to test this point, we had to rule out the possibility that infants might detect face-sound correspondences that were based purely on temporal grounds. Thus, the two visual images had to be presented in perfect temporal synchrony with one another. The two mouths had to open and close at the same time. Moreover, the sound track had to be aligned with the filmed images so that the sound was temporally synchronous, and equally so with both the "matched" and the "mismatched" mouth movements (Kuhl, in press, a).

Filming and selecting temporally matched stimuli. A film was made in a studio of a female talker producing the vowels /a/ (as in "pop") and /i/ (as in "peep"). The talker placed her head through an oval hole in a black velvet cloth so that only her face was filmed. A 16-mm camera (Arriflex BL) recorded the image on color film. Audio recordings were made on a high-quality tape recorder (Nagra, IV-L) and transferred to 16-mm magnetic sound track.

The talker produced the vowels once every 3 s and attempted to articulate them with equal intensity and duration. Rather than using a single production of /a/ and a single production of /i/ and trying to match them on all noncritical dimensions (visual extent of mouth movement, visual rate of mouth movement, auditory intensity, auditory duration, and fundamental frequency), we used a number of different productions to represent the /a/ and /i/ categories. The stimuli selected for use in the experiment were chosen so that the noncritical dimensions fell within a narrowly restricted range that overlapped for the two vowel categories. This procedure helped ensure that recognition of a correspondence between face and sound could not be based on an idiosyncratic property of a single production of the vowels.

Two sequences of 10 /a/'s and two sequences of 10 /i/'s were chosen as stimuli from among the filmed images. They were used to make two film loops. One loop displayed the /a/ face on the right and the /i/ face on the left; the second loop reversed this left-right orientation. The facial images were chosen such that the durations of the individual articulations fell within a narrowly defined range that overlapped for the two vowel categories. The duration of each visual stimulus was measured to specify the length of time that the lips were parted. The average duration of /a/ mouth movements was 1.92 s (range, 1.75–2.08 s). The average duration of /i/ mouth movements was 1.97 s (range, 1.79–2.13 s).

A comparable process was used to choose the sequences of /a/'s and /i/'s from the audio recording. The auditory stimuli used in the experiment were not those emanating from the individual articulations shown on the visual loop. This further ensured that no idiosyncratic characteristics of a particular production would link the auditory and visual displays. (For example, a visual lip twitch that corresponded to a slight jitter in the fundamental frequency might otherwise cue the correct pairing.) A storage oscilloscope (Tektronix, Model 7603) was used to measure the duration of each of the auditory stimuli. Duration was specified in ms, starting with the first pitch period and ending with the last pitch period that demonstrated the periodicity characteristic of these vowels. The average duration of the /a/ vowels was 1.15 s (range, 1.06–1.27). The average duration of the /i/ vowels was 1.12 s (range, 1.05–1.22). Note that the auditory stimuli were shorter than the visual stimuli. This is due to the fact that in natural speech sound is not emitted the instant that the lips part, and audible sound does not continue until the instant that the lips come together again.

Spectrographic measurements were made of the formant frequencies of the /a/ and /i/ stimuli. The average frequencies of the first three formants for the /a/ stimuli were 741 Hz, 1065 Hz, and 3060 Hz. The average frequencies of the first three formants of the /i/ stimuli were 416 Hz, 2338 Hz, and 2718 Hz. The /a/ and /i/ vowels were produced with a rise-fall fundamental frequency contour. These contours were characterized by an initial rapid rise in frequency over the first 200 ms, followed by a longer, more gradual decline in the fundamental over the remainder of the vowel. The average starting frequency of the contours was 204 Hz, the average peak 276 Hz, and the average final frequency 160 Hz.

Synchronizing the stimuli. Once the auditory and visual stimuli were chosen, the two facial images had to be synchronized with each other and combined to produce a two-image film loop. Then the sound tracks had to be synchronized to the two-image film track. Using a studio-quality editing bench, the facial images were aligned so as to minimize any differences between the onsets and offsets of the /a/ and /i/ gestures. The alignment was accomplished so that the average difference between the onsets and offsets of each /a-i/ pairing was less than one frame (0.042 s). Thus the visual stimuli were assembled to ensure that the /a/ mouth movements and the /i/ mouth movements were in visual synchrony with one another.

The synchronization of the film and sound tracks was also done using the editing bench. We examined a variety of sync points, looking for the one that appeared to be the most natural. Since the /a/ and /i/ visual stimuli were equally long and had equivalent start points, the alignment of each sound track (whether /a/ sounds or /i/ sounds) was equally good, from a temporal standpoint, to both faces. In sum, there were no obvious temporal cues linking the /a/ vowels specifically to the /a/ faces, rather than the /i/ faces, and vice versa.

When projected, the faces were approximately life-sized, 21 cm long and 15 cm wide. Their centers were separated by 38 cm. The sounds were presented at an average intensity of 60 dB SPL (range for /a/'s, 56–64; range for /i/'s, 55–62).

Equipment and Test Apparatus

Figure 1 shows the test suite. It consisted of two sound-treated rooms separated by clear glass. The control room housed the projector, a DEC PDP 11-34 computer, and the video recorder and monitor. The film was projected through the glass into the experimental room.

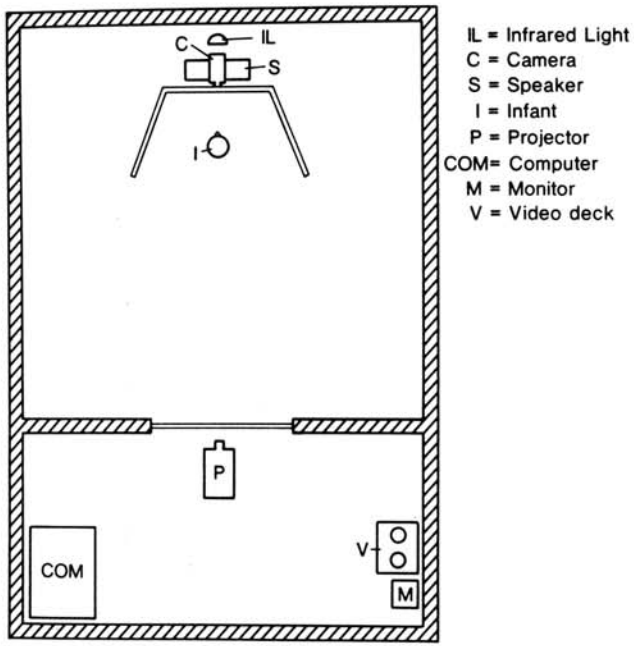

Figure 1. Schematic of the test suite.

The stimuli were projected using a dual system projector (Siemens, Model 2000). This projector incorporates an interlocked sprocket assembly for driving the film and sound loops. Thus the 16 mm film and sound tracks could not get out of synchrony once they were aligned and started. The film and sound tracks contained 10 articulations, measuring 3 s from stimulus onset to stimulus onset. Specially designed loop holders were constructed to allow the film and sound loops to recycle for the duration of the experiment.

The infants were tested facing a three-sided cubicle in the experimental room. The cubicle consisted of three white panels: a front panel on which the films were projected and two side panels positioned at 75° angles from the front panel. Infants were tested while in an infant seat placed on a table. When

seated, they were 46 cm from the facial displays. The loudspeaker (Electrovoice SP-12) was located behind the front panel and centered midway between the two facial images.

During testing, the major source of visible light was that provided by the films. An infrared lamp was suspended above the test cubicle. An infrared-sensitive camera (Panasonic, Model WV-1354A) was positioned behind a small hole located between the two faces. It recorded a close-up of the infant's face on the video recorder (Sony 3650). Observers used these videotaped records to score the infant's fixations to the right and left facial images at the completion of the experiment. A microphone, which recorded infants' vocalizations, was suspended above the test cubicle.

Procedure

The experimental procedure involved two phases, a familiarization and a test phase. During the familiarization phase each visual stimulus (the */a/* face and */i/* face) was presented without sound for 10 s. This was accomplished by occluding one facial image and then the other, for 10 s each. After this 20 s period, both visual stimuli were briefly occluded until the infant looked to midline. Then, the sound (either */a/* or */i/*) was turned on and both faces were presented for the 2-min test phase. The sound presented to the infant, the left-right positioning of the two faces, the order of familiarization, and the sex of the infants were counterbalanced.

Scoring

Infants' visual fixations were scored from videotape by an independent observer. The videotape showed only the infant's face. The observer could neither hear the sound nor see the faces presented to the infants, and thus was appropriately unbiased. The observer pressed buttons on a Rustrack event recorder to indicate when the infant was looking at the left or right visual display. Reliability, both intra- and interobserver, was assessed by rescoring the entire data set of 32 subjects. The results showed high scoring agreement. Using the percentage of total fixation time recorded to the matching face as a measure, the mean difference between the intraobserver assessments was 1.8%. For the interobserver assessment, the comparable score was 3.3%. As assessed by the Pearson *r*, the intra- and interobserver agreements were both 0.99.

RESULTS

Of the total test time, infants spent 90.3% looking at one or the other of the two faces. The percentage of total fixation time devoted to the face that matched versus the one that mismatched the auditory stimulus was calculated for each infant. The mean percentage devoted to the matched face was 73.6%. This is significantly greater than the 50% chance level, $t(31)=4.67$, $p<0.001$. Of the 32 infants, 24 looked longer at the matched face. This is significant by the binomial test ($p<0.01$).

Other factors counterbalanced in the design were also examined statistically. None proved significant: Infants did not look significantly more to the right as opposed to the left side, nor did they prefer to look at the */a/* face as opposed to the */i/* face, or to the face seen last during familiarization. Specifically, the mean percentage of total fixation time devoted to the right side was 46.1%, $t(31)=0.60$, $p>0.50$; the mean percentage of total fixation devoted to the */a/* face was 46.5%, $t(31)=0.53$, $p>0.50$; and the mean percentage of total fixation devoted to the face seen last during familiarization was 50.9%, $t(31)=0.13$, $p>0.50$. Finally, there were no differences in the percentage of fixation time devoted to the matched face when it appeared on the right side as opposed to the left side, $t(30)=0.078$, $p>0.40$.

DISCUSSION

The experiment demonstrated that the auditorially presented vowel influenced infants' visual fixations. Infants looked longer at the face that matched the sound. When listening to the */a/* vowel, infants looked longer at the */a/* face; when listening to the */i/* vowel, infants looked longer at the */i/* face. This shows that infants recognize a cross-modal correspondence between auditorially and visually presented speech information.

What is the basis for this cross-modal effect? We want to claim that infants recognize the correspondence between particular articulatory gestures and the particular sounds they cause. This idea implies that infants recognize correspondences between certain properties of the visually-specified vowels, such as the lip shapes, jaw positions, etc., and the formant frequency patterns of the auditory stimuli.

This claim cannot be advanced unless the temporal hypothesis is ruled out. The temporal account argues that the face-voice pairs are linked by temporal cues, such as the detection of a match between the onset of acoustic energy and the parting of the lips. If the cross-modal matching effect were attributable to the temporal properties of the sound rather than to the spectral properties of the sound, one could not claim that infants relate */a/* faces to */a/* sounds, only that they relate parting lips to sound energy.

The elaborate procedure used to synchronize the auditory and visual stimuli in Experiment 1 was, in fact, aimed at eliminating the temporal hypothesis. Nevertheless, we wanted to test the hypothesis directly in Experiment 2. We altered the original auditory stimuli used in Experiment 1 so as to remove the spectral information from the vowels while preserving their temporal characteristics. The stimuli used in Experiment 2 preserved the on/off, durational, and amplitude-envelope characteristics of the original vowels. Thus, if temporal cues alone had allowed infants to link the auditory and visual stimuli in Experiment 1, the cross-modal matching effect should still obtain.

While we expected infants' performance to drop to chance in Experiment 2 (because we hypothesized that the temporal cues were not sufficient to reproduce the effect), we did not consider the test trivial. Relatively small asynchronies between auditory and visual events may be detectable. There are no

data on which to base guesses about how small a discrepancy might be sufficient to cue the effect. Adult data showing that it takes a 131-ms discrepancy between auditory and visual events to identify asynchrony (Dixon & Spitz, 1980) are not pertinent, because the task demands in our test were different. Dixon and Spitz's task required that a single auditory-visual pairing be classified as "out of sync" by observers, whereas the type of paired-comparison task used here simply requires that one of the two faces is perceived to be in "better" temporal alignment with the sound than the other. The available data on infants' detection of asynchrony between auditory and visual events are also not helpful. Studies have suggested that infants may be able to detect very gross temporal asynchronies (400 ms) between mouth movements and sound (Dodd, 1979), but they have not tested infants' ability to detect slight asynchronies of the type at issue here. Until experiments on infants' detection of slight temporal asynchronies have been done, the temporal hypothesis must be ruled out by demonstrating experimentally, in each specific instance, that infants cannot solve the cross-modal matching problem when only temporal information remains.

Moreover, the signals we used in Experiment 2 can serve to test another explanation that can complicate the interpretation of auditory-visual speech-perception tests. A recent test of auditory-visual speech perception by infants may be particularly subject to this problem (MacKain, Studdert-Kennedy, Spieker, & Stern, 1983). In that experiment, infants were given 12 different trials, each 20 s in duration. Six different CVCV syllables formed pairs, such as "mama" versus "lulu" and "baby" versus "zuzi." These pairs were visually presented while a sound matching one of the two syllables was presented auditorially.

Two points need to be made about the study. First, the overall effect of looking to the matched face was significant, and thus our original cross-modal matching effect for speech was replicated. At the same time it should be noted, however, that the outcome was somewhat weak and complicated by many factors, including the fact that the effect was not produced for any of the six disyllables when the matched stimulus was located to the infant's left, and only three of the six disyllables when it was located to the infant's right. The weakness of the effect may have been due to the large number of conditions and the repeated-measures design; infants may be biased to look back to the side that resulted in a "match" on the last trial, particularly when it occurred on the infant's right.

The second and more important point, however, is that their effect is particularly subject to a form of the temporal hypothesis outlined earlier. The MacKain et al. experiment involved disyllables whose auditory and visual concomitants could be related simply by detecting a correspondence between the amplitude envelope of the sound and the degree of mouth opening. The argument is as follows. Syllables involving stop consonants, such as "baby," involve a complete occlusion of the vocal tract at the juncture between the two

syllables. Syllables that do not contain stop consonants, such as "lulu," do not involve an occlusion of the tract. This means that when "baby" and "lulu" are visually paired infants will see a complete closure of the lips during "baby" but very little lip movement during "lulu." The sounds corresponding to "baby" and "lulu" show corresponding changes in their amplitude envelopes. Sound ceases momentarily between the two syllables of "baby." For "lulu," the amplitude decreases only slightly at the juncture between the two syllables, producing an almost continuous amplitude envelope. Thus, infants could solve this cross-modal matching problem by recognizing the correspondence between lip closure and the cessation of sound, or continued lip opening and the continuance of sound. This would not be an uninteresting finding, but it could not support our claim that infants recognize the correspondence between lip shapes and formant frequencies—only that they recognize general synchronies between the amount of change in visual movement and the amount of change in auditory amplitude.

One way of assessing this temporal explanation of the cross-modal speech effect is to design an experiment in which all spectral cues are eliminated, and only timing and amplitude-envelope ones remain. This was the basic rationale for Experiment 2. It isolated the temporal information available in Experiment 1 to see if it was sufficient to reproduce the effect. Demonstrating that this information is insufficient to reproduce the cross-modal matching effect is critical to advancing the hypothesis that it is speech information *per se* that is cross-modally represented.

EXPERIMENT 2

METHODS

Subjects

A new group of 32 normal, full-term infants was tested. The infants ranged in age from 18.1 to 20.0 weeks ($M = 19.4$). An additional 12 infants failed to complete testing due to crying (4), falling asleep (5), or equipment failure (3). Their participation was solicited in the same way as that previously described and the selection criteria were identical.

Stimuli and Equipment

Experiment 2 used the same visual stimuli as previously. The auditory stimuli were altered to remove the spectral information necessary to identify the vowels while preserving certain temporal characteristics. The stimuli consisted of pure tones centered at 200 Hz, the average of the female talker's fundamental frequency. Each of the original vowels was replaced by a pure-tone stimulus that preserved three temporal aspects of the original vowel: (a) its durational characteristics; (b) its amplitude envelope over time; and (c) its alignment in relation to the visual stimuli.

In order to reproduce the precise alignment between the visual stimuli and these altered auditory stimuli, the output of the original 16 mm sound track was used to trigger the presentation of the sine-wave envelope analogs by the computer. This occurred virtually instantaneously; the onsets of the computer-generated analogs occurred within 40 microseconds of the original vowels. The equipment and apparatus used in Experiment 2 were identical to those described previously.

In sum, infants in Experiment 2 were presented with the same two faces articulating */a/* and */i/*, but heard pure-tone stimuli instead of vowels. These stimuli followed the duration, amplitude envelope, and onset/offset characteristics of the original vowels. The tone stimuli became louder as the two mouths opened wider and diminished in intensity as the two mouths began to close, just as the vowels had. If infants had relied on temporal cues to solve the cross-modal task in Experiment 1, then the cross-modal effect should obtain in this experiment.

Procedure

All aspects of the procedure were identical to those described in Experiment 1. Intra- and interobserver agreement, assessed as before, was again high. Using the percentage of total fixation time recorded to the matching face as the measure, the mean difference between the two intraobserver assessments was 1.6%; the mean difference between the interobserver assessments was 4.9%. As assessed by the Pearson *r*, the intra- and interobserver agreements were both 0.99.

RESULTS AND DISCUSSION

Of the total test time, infants spent 93.1% looking at one or the other of the two faces. The percentage of total fixation time devoted to the "matched" versus the "mismatched" face was calculated for each infant. The mean percentage fixation to the matched face did not differ significantly from chance ($M = 54.6\%$; $t(31) = 0.78$, $p > 0.50$). Of the 32 infants, 17 looked longer at the matched face, and 15 looked longer at the mismatched face.

The fixation time scores were also analyzed to determine if any of the additional main factors produced significant differences. In Experiment 1, none were significant; in this experiment, two of the three were not. The two nonsignificant factors were: the percentage of total fixation time spent looking at the right side ($M = 53.9\%$; $t(31) = 0.66$, > 0.50), and at the last familiarization side ($M = 51.0\%$; $t(31) = 0.17$, $p > 0.50$). The */a-i/* face comparison reached significance, however. Infants spent a greater percentage of the total fixation time looking at the */a/* face, rather than at the */i/* face ($M = 62.7\%$; $t(31) = 2.28$, $p < 0.05$). The same visual stimuli were presented in both Experiments 1 and 2, and no face preferences appeared in Experiment 1; perhaps infants tended to fixate the face they most preferred in the absence of a cross-modal correspondence between the auditory and visual stimuli.

GENERAL DISCUSSION

We examined infants' abilities to detect auditory-visual correspondence for speech in two experiments. In Experiment 1 infants were shown two faces, articulating in synchrony, two different vowels. They were auditorially presented with a single vowel that matched one of the two faces. The results demonstrated that the auditory stimulus systematically influenced infants' visual preferences. Infants looked longer at the matched face rather than at the mismatched face. Thus, infants detect a correspondence between the auditory and visual concomitants of speech.

In Experiment 2 infants were presented with the same visual stimuli, but with auditory stimuli that were altered. The altered stimuli were pure tones—the spectral cues that are necessary and sufficient to identify the vowels were removed. The altered stimuli did, however, preserve certain temporal features of the original vowels. Specifically, they maintained the onset-offset characteristics, durations, and amplitude envelopes. Moreover, they maintained the exact temporal correspondence between the auditory and visual stimuli that was present in Experiment 1.

If infants were relying on temporal information to link a particular articulatory gesture to a particular sound in Experiment 1, the cross-modal matching effect should also have been obtained in Experiment 2. Performance, however, dropped to chance. Apparently, the temporal information contained in the waveform envelopes of the auditory stimuli was not sufficient, in the absence of spectral information, to produce the matching effect. Infants did not link sounds to faces on a purely temporal basis; therefore, some aspect of the spectral information was shown to be critical to the detection of these correspondences. This latter finding was in accord with our predictions, given the elaborate procedure used to align the auditory and visual stimuli in Experiment 1.

THE INTERMODAL REPRESENTATION OF SPEECH: ITS DEVELOPMENT AND BASIS

These experiments demonstrate that 4.5-month-old infants relate the auditory and visual products of articulation. The findings suggest that information about speech is intermodally represented in infants. Two important questions emerge from these results. The first pertains to the development of the ability to detect auditory-visual concomitants for speech. The second pertains to its basis.

Development of the Effect

There are three interesting accounts of the development of this cross-modal ability. First, young infants have had experience watching caretakers speak and may have learned specific auditory-visual pairings. A particular articulatory posture (an open mouth) might have become associated with the sound */a/*; another (spread lips) with the sound */i/*. By this account, the events in the

two modalities would simply have been associated while listening to and watching talkers speak.

The extent to which this account works as an explanation for the detection of cross-modal correspondences for speech can be tested either by using younger infants, or by examining the effect with speech sounds the infant has never seen—that is, ones not occurring in their native language. If infants succeed on tests involving sounds they have neither seen nor heard, it would go against this simple learning account.

A second, more complex developmental account could also be offered to explain infants' abilities to detect auditory-visual correspondences for speech. This one also holds that experience is essential in producing the effect. In this case, however, the argument would be that it is experience in producing speech sounds oneself that is the prerequisite for detecting auditory-visual correspondences for speech. Since the infants in our experiment were 4.5 months old, they were well into the "babbling phase" (Oller, 1980). They had almost certainly produced one of the vowels used in the experiment (the vowel */a/*, but probably not the vowel */i/*; Lieberman, 1980).

How would the "babbling account" work?[1] In order to explain how babbling and the cross-modal perception of speech are tied, three ideas have to be developed. The first is that our auditory-visual test can be viewed as posing an auditory-articulatory mapping problem to infants. The second is that infants may acquire knowledge of the relationship between audition and articulation during babbling. The third is that in order for infants to use the knowledge gained during babbling (when they relate their own mouth movements to sound) to solve our cross-modal test (which requires that infants relate another talker's mouth movements to sound), additional requirements must be met. Let us turn now to the first of these three.

Thus far, we have viewed this experiment as a test of auditory-*visual* perception. Now we take note of the specific nature of the visual information presented in this experiment—the movements of the articulators. There are many ways to present speech information to the eye. But in this case, and in all those involving lipreading, the information is of a particular kind. Lipreading does not simply consist of a visual transform of speech, such as that provided by a visual on-line record of the amplitude or the formant frequencies. Lipreading involves real articulators—lips, tongue, teeth, and jaw—moving to form particular configurations.

To the extent that the articulatory movements are visible, watching them provides direct information about the speech movements that actually occurred. It is the only sure way of determining what the articulators did, since no other measure relates directly to articulation. Other displays, such as the on-line visual record of some acoustic measurement, or even the auditory percept itself, do not uniquely specify the articulatory movements. While the articulatory information obtained by eye is limited (because not all of the relevant movements can be seen), the information is direct and unambiguous. Thus, watching a talker's lips come together provides direct and unmistakable evidence that a bilabial articulation occurred.

[1] To facilitate the description of this theoretical account, we have referred to it as the "babbling account," even though infants at this age (4 months) are more engaged in "cooing" or "sound play" than they are in producing the reduplicated CV sequences that are characteristic of infants at an older age. This latter type of speech has now become the classic example of "babbling."

The implication for this discussion is that while we have conducted an auditory-*visual* experiment, we have employed a visual transform that involves *articulation*. This means that the experiment poses an auditory-articulatory mapping problem to infants, one in which they must recognize the correspondences between articulatory gestures presented visually and the sounds that emanate from mouths moving in that way.

How does this impact the babbling account? Explaining how it does involves our second point, namely that the experience gained during babbling may enable infants to recognize auditory–articulatory correlates. Theoretical interpretations of babbling hold that it is during this period infants are mapping out the enormously complex relationship between articulatory maneuvers and their auditory results (Netsell, 1981; Oller, 1980). It is at this time that infants may learn that a mouth-open/tongue-low posture results in a sound like */a/*, while a lips-spread/tongue-high posture results in a sound like */i/*. Presumably, the eventual mastery of a set of rules concerning auditory–articulatory mapping allows infants to produce a specific sound, at will, with their own articulatory mechanisms. But there are no data clearly demonstrating that prebabbling infants fail to understand the relationship between sound and articulatory movement and that they gain this knowledge during the babbling period. One of the problems is, of course, designing a measure of infants' knowledge of auditory–articulatory relations that could be used to test the hypothesis.

The cross-modal speech task used in this experiment provides one approach to doing this. It poses the auditory–articulatory problem to infants by asking whether they recognize the correspondence between articulatory gestures and sounds when they are produced by a talker other than themselves. Developmental studies could show that prebabbling infants fail on our cross-modal task, and that they succeed only after they have mastered a set of rules for relating audition and articulation; in other words, only after the babbling phase is underway.

The third and final point related to the babbling account is that in order for infants to use their babbling experience to solve our cross-modal matching task, two requirements must be met: Infants must recognize the equivalence between the articulatory actions they see another produce (the visual stimuli in our experiment) and the articulatory actions they themselves can produce. Moreover, they must recognize that the vowel sounds produced by another (the auditory stimuli in our experiment), are equivalent to ones they themselves can produce. Without these abilities, any knowledge of auditory–articulatory relations infants gained during the babbling phase, when they were producing

speech themselves, could not be used in the recognition of correspondences between articulatory actions and sounds produced by a talker other than themselves.

Studies on the imitation of oral movements presented visually to infants and studies on speech-sound categorization in infants provide evidence suggesting that both prerequisites are at least within the capabilities of young infants. Regarding the first of these, studies show that infants are capable of imitating facial gestures such as a mouth opening/closing in the first weeks of life (Meltzoff & Moore, 1977) and even the first days (Meltzoff & Moore, 1983a). These results suggest the existence of a basic mechanism that enables infants to relate oral movements of their own to equivalent movements they see another produce (Meltzoff, in press; Meltzoff & Moore, 1983b).

Regarding the second, studies show that infants at this age categorize as similar identical vowels produced by men, women, and children, even though the vowels are highly discriminable and very different acoustically (Kuhl, 1979; 1983). Thus, young infants preserve an auditory constancy for identical vowels produced by different talkers. It is therefore likely that they equate the vowels produced by others and those they themselves produce (Kuhl, in press, b).

Without these perceived equivalences—between actions they see and ones they produce, and between sounds they hear and sounds they produce—infants would be unable to use the experience gained through babbling to solve the cross-modal matching problem. The combination of the auditory-articulatory links acquired in babbling with mechanisms that allow them to equate actions they see with ones they themselves produce, and sounds they hear with ones they produce, may well underlie infants' detection of cross-modal concomitants for speech. If so, babbling could be a developmental precursor to the detection of cross-modal correspondences for speech.

A third developmental account is that knowledge of the relationship between particular articulatory gestures and their auditory results is innate—that is, that infants are born with the ability to relate the auditory and visual products of articulation. This means that prior to experiencing auditory-visual correspondences in the speech of others and prior to babbling, infants would be capable of detecting these cross-modal correspondences. Experiments of the type reported here would have to be conducted with newborns in order to provide definitive support for this claim.

The Basis of the Effect

The results of Experiment 1 constitute the first evidence that infants recognize correspondences between auditory and visual information for speech. One fundamental question is: What is the basis of the perceived match between the auditory and visual information? Is speech itself or something else intermodally represented?

The results of Experiment 2 provide information concerning the basis of the effect. The results show that infants' detection of auditory-visual correspondences for speech depends on some aspect of the spectral information contained in the sounds. This is an important finding because it suggests that infants relied on something other than temporal information to link the optic and acoustic events. Had it been the case that they simply used gross timing or amplitude information, nothing could be claimed about infants' detection of the relationship between sound patterns and particular articulatory gestures, and the possibility of an intermodal representation of phonetic information would not even be raised. The fact that temporal information was not sufficient, however, suggests that infants relate spectral properties of sounds to particular articulatory postures.

Given that the effect depends on spectral information, there are still two distinct possibilities. One hypothesis is that the effect relies on the recognition of phonetic information in the signals picked up through the two modalities. This "phonetic account" can take one of two forms. The first is an "identity match," wherein the same phonetic information is specified by the two different modalities. The phonetic information could be matched either at the whole-unit level (phone) or at the feature level. For example, if the visual information specifies a phone such as */i/*, and the auditory information also specifies */i/*, then a cross-modal match could be based on the recognition of a common phonetic identity at the level of the phone. But a feature-level match also bears consideration. Auditory and visual speech information would often need to be matched at the level of the feature, rather than at the level of the phone, due to the imprecision with which the eye can deliver the information. For example, when the visual stimulus is the syllable */ba/*, the optic stream cannot report */ba/*, rather than */pa/* or */ma/*, but only that the sound is a "bilabial." Thus, if the bilabial feature is represented in both modalities, the matching effect could be based on the recognition of a common phonetic identity at the level of the feature. In both cases, regardless of whether the match is at the level of the phone or the feature, the information represented auditorially and visually is the same.

The "phonetic account" also offers a second possibility, one that does not involve an identity match. The second possibility is that the inputs to the two modalities, while not specifying the identical information, are related because they are tied to a common phonetic representation. In this case, the match is based on the fact that the information entering the two modalities is tied at a higher level of representation, one that unites the information that is delivered to each of the two modalities. This version of the account would thus involve a "supramodal" phonetic representation. We know from lipreading experiments that a phonetic percept can result from the combination of nonidentical featural information in the two modalities. When visual information conveying the place feature "bilabial" is combined with auditory information conveying the manner feature "voiced," the percept */ba/* results. The */ba/* was not presented to either modality alone—it is a combination of information picked up by ear and by eye. Thus, perceived cross-modal matches could be

based on the fact that the auditory and visual inputs, while not conveying identical information, are tied to a common, "supramodal" phonetic representation.

The important point here is that both versions of the "phonetic account" imply that infants have access to a representation of phonetic units that specifies both their auditory and visual realizations. Input from the two modalities —optic and acoustic—is linked by these phonetic representations. Thus, the phonetic account outlined above argues that it is speech *per se* that is intermodally represented.

A second, very different, hypothesis may also be entertained—namely, that the basis of the perceived match is independent of speech. This "wholly independent of speech" account reflects a logical alternative, but one that is difficult to prove under most circumstances. Consider the simplest case. If correspondences between auditory and visual events are based on cues that do not contribute to phonetic identity, then it can be concluded that the auditory and visual events are related wholly independently of speech. A good case is one in which amplitude information relates the two streams. In most instances (but not all) phonetic units are not defined in terms of amplitude, so if the cross-modal correspondence is based on these cues, it is done independently of speech.

Now consider a more complex case, one in which the cues relating the two streams are those used to identify the sounds. For vowels in this experiment this involves spectral information. The spectral features of vowels include "compact" and "grave" (for the vowel */a/*) and "diffuse" and "acute" (for */i/*) (Jakobson, Fant, & Halle, 1965). These features can be isolated and conveyed in a nonspeech context. This means that certain nonspeech sounds mimicking the spectral properties of speech, without being identified as speech, might produce the cross-modal effect. If so, what would we conclude? We could then say that the cross-modal matching effect is not *restricted to speech* (because nonspeech sounds are sufficient to produce it), but we would not want to argue that the effect is *wholly independent* of speech because it may indeed be the case that the effect obtains precisely because the nonspeech sounds capture the critical properties of speech (Kuhl, in press, c).

The strongest test of the "wholly independent of speech" account would be if the identical information lead to two distinct and separate predictions, one based on a phonetic interpretation of the information, and the other based on a purely physical interpretation of the information. This would be the case if a match between the information in the two modalities based on pure psychophysics would lead to one outcome, while a match based on phonetics lead to the opposite outcome. If such a test could be devised, it would help determine whether infants' cross-modal matches for speech were based on "physics" or "phonetics."

A final point related to the basis of the effect should be made. To what extent does finding that speech information is intermodally represented in infants afford it a unique status? Recent work shows that the detection of intermodal correspondences is not restricted to speech. Rather, it appears that young infants have a general ability to detect and utilize intermodal equivalences in the information picked up by different modalities, and that this ability underlies their success on a variety of cross-modal tasks (Bower, Broughton, & Moore, 1970a, b; Meltzoff & Borton, 1979; Meltzoff & Moore, 1977; 1983a; Spelke, 1979). The present experiment extends this previous work by providing another example of young infants' cross-modal abilities. The degree to which these speech effects uniquely differ from, or are a special case of, more general intermodal abilities is yet to be determined. Nonetheless, the recognition of cross-modal correspondences for speech promises to provide a particularly rich and detailed set of examples with which to examine the nature and bases of infants' abilities to relate information picked up through different modalities.

In summary, the present experiments show that by 4.5 months of age infants detect a relationship between the auditory and visual concomitants of speech. They recognize that the sound */a/* emanates from a wide-open mouth while the sound */i/* emanates from a mouth with spread lips. Whether future research will reveal that this knowledge derives from an intermodal representation of phonetic information or whether it stems from a match between the properties of sounds and the properties of lips independent of speech does not detract from its relevance to linguistic development. The fact that infants recognize that particular sounds correspond to mouths moving in particular ways suggests that speech itself, or the information underlying speech perception, is intermodally represented in young infants.

REFERENCES

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of selected English consonants. *Journal of Speech and Hearing Research, 17,* 619–630.

Bower, T. G. R., Broughton, J. M., & Moore, M. K. (1970a). Demonstration of intention in the reaching behaviour of neonate humans. *Nature, 228,* 679–680.

Bower, T. G. R., Broughton, J. M., & Moore, M. K. (1970b). The coordination of visual and tactual input in infants. *Perception and Psychophysics, 8,* 51–53.

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9,* 719–721.

Dodd, B. (1979). Lipreading in infants: Attention to speech presented in- and out-of-synchrony. (1979). *Cognitive Psychology, 11,* 478–484.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research, 12,* 423–425.

Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research, 15,* 413–422.

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders, 40,* 481–492.

Erber, N. P. (1977). Speech perception by profoundly deaf children. In J. M. Pickett (Ed.), Papers from the Research Conference on Speech Processing Aids for the Deaf, Gallaudet College, Washington, DC, 2–19.

Ewertsen, H. W., & Birk-Nielsen, H. (1971). A comparative analysis of the audio visual, auditive, and visual perception of speech. *Acta Otolaryngology, 72,* 201–205.

Grant, K. W., Ardell, L., Kuhl, P. K., & Sparks, D. W. (in press). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to connected discourse perception by speech readers. *Journal of the Acoustical Society of America.*

Jakobson, R., Fant, C. G. M., & Halle, M. (1965). *Preliminaries to speech analysis: The distinctive features and their correlates.* Cambridge, MA: M.I.T. Press.

Johnson, S. (1775). *A journey to the western islands of Scotland.* London: Simpkin, Marshall, Hamilton, & Kent.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America, 66,* 1668–1679.

Kuhl, P. K., (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development, 6,* 263–285.

Kuhl, P. K. (in press, a). Methods in the study of infant speech perception. In G. Gottlieb and N. Krasnegor (Eds.), *Measurement of audition and vision in the first year of postnatal life: A methodological overview.* Norwood, NJ: Ablex.

Kuhl, P. K. (in press, b). Categorization of speech by infants. In J. Mehler and R. Fox (Eds.), *Neonate cognition: Beyond the blooming, buzzing confusion.* Hillsdale, NJ: Erlbaum.

Kuhl, P. K. (in press, c). Reflections on infants' perception and representation of speech. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes.* Hillsdale, NJ: Erlbaum.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218,* 1138–1141.

Lieberman, P. (1980). On the development of vowel production in young children. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.), *Child phonology: Vol. I. Production.* New York: Academic Press.

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science, 219,* 1347–1349.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Meltzoff, A. N. (in press). The roots of social and cognitive development: Models of man's original nature. In T. M. Field & N. Fox (Eds.), *Social perception in infants.* Norwood, NJ: Ablex.

Meltzoff, A. N., & Borton, R. W. (1979). Intermodal matching by human neonates. *Nature, 282,* 403–404.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198,* 75–78.

Meltzoff, A. N., & Moore, M. K. (1983a). Newborn infants imitate adult facial gestures. *Child Development, 54,* 702–709.

Meltzoff, A. N., & Moore, M. K. (1983b). The origins of imitation in infancy: Paradigm, phenomena, and theories. In L. P. Lipsitt (Ed.), *Advances in infancy research* (Vol. 2). Norwood, NJ: Ablex.

Miller, G. A., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338–352.

Netsell, R. W. (1981). The acquisition of speech motor control: A perspective with directions for research. In R. E. Stark (Ed.), *Language behavior in infancy and early childhood.* New York: Elsevier North Holland, Inc.

Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. Yeni-Komshian, C. A. Ferguson, & J. Kavanagh (Eds.), *Child phonology: Vol. I. Production.* New York: Academic Press.

O'Neill, J. J. (1954). Contributions of the visual components of oral symbols to speech comprehension. *Journal of Speech and Hearing Disorders, 19,* 429–439.

Sanders, D. A., & Goodrich, S. J. (1971). The relative contribution of visual and auditory components of speech to speech intelligibility as a function of three conditions of frequency distortion. *Journal of Speech and Hearing Research, 14,* 154–159.

Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology, 15,* 626–636.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica, 36,* 314–331.

15 June 1983; Revised 5 January 1984 ■